European Scientific Journal October 2014 edition vol.10, No.29 ISSN: 1857 – 7881 (Print) e - ISSN 1857- 7431

THEORETICAL AND PRACTICAL PROBLEMS IN DESIGNING TESTS FOR FOREIGN LANGUAGE PROGRAMS

Leonard Rapi
Harallamb Miconi
Department of Foreign Languages,
University "Eqrem Cabej", Gjirokaster, Albania

Abstract

Testing is an important component in foreign language programs. This occurs for a number of reasons. First, foreign language teachers systematically test their students in order to assess their performance and achievements in the learning process. On the basis of their test results, teachers make decisions which affect their students' further progress. For example, they decide whether their students have acquired the material for which they have been tested or whether action should be taken to help those whose performance results problematic. In more extreme cases, it is test results again that determine whether a student may move on to an upper level or not. Second, tests may have scientific purposes. They may be applied to foreign language classes as part of scientific studies to collect data linked to various aspects of learning-teaching foreign languages. In this context, an issue of great importance is that of the reliability and validity of the administered tests. To what extent can we claim that these tests serve as reliable indicators of the level of acquisition of the subject material by our students? The fact is that when there is practically no training whatsoever in the field of test development, teachers often rely mostly on their own intuition or their previous experience as students or turn to tests that come with the textbooks they use. In view of the above, this article will discuss some problems which have to do with test design and development in language programs as well as their possible implications for our teachers.

Keywords: Reliability, construct validity, threats to reliability, threats to validity

Introduction

In this paper a number of issues are discussed which have a direct bearing on tests and testing in foreign language programs. More specifically,

we focus on the use of tests in language programs, as well as other relevant issues such as test reliability and test validity. Based on the literature on this field, we first try to build a case for the usefulness of tests in language courses. Next we concentrate on test reliability and test validity and various related problems. All along we discuss the potential implications these problems could have for the language teacher as well as offer suggestions as to what teachers can do to design reliable and valid tests that could better serve their purposes.

Use of Tests in Language Programs

The primary use of tests in language programs is that of providing feedback about students' level of learned language abilities (Henning, 1987; Brown, 1996; Bachman, 1996). This means that tests are used to help identify strong and weak points in students' language skills. For example, through tests teachers can see that their students seem to be doing quite well in grammar or reading comprehension but that their speaking or listening abilities are not what they are expected to be. These tests are usually called *diagnostic* or *achievement* tests because they give valuable information that can help students, teachers or school administrators to make proper decisions in order to make language programs more effective.

Tests called *proficiency tests* are often used for selection purposes. (Brown, 1996) Based on test scores, decisions are made to select who can be accepted in a particular program of instruction and who can't on the basis of language criteria. Examples of these tests are standardized tests such as TOEFL, FCE, IELTS etc. Students need to pass these tests at a certain score level in order to meet the language requirements set by universities to be allowed to participate in the programs for which they have applied.

There have been claims that due to their biases these tests often penalize test takers in a number of ways and their fairness has often been disputed (Henning, 1987). We tend to agree with this and we can offer a specific example based on our observation because for the moment we do not have empirical data to support our claim. We suspect that Albanian students who take IBT TOEFL are in a way unjustly penalized because of the fact that TOEFL is administered totally online. This may happen for two main reasons. First, the internet service is less than reliable. It can be slow and defects happen a lot. Second, due to lack of computer and typing skills our students have considerable difficulty following the test procedures and very often they are not able to complete all the tasks in time. This seems to be a real problem for our students at this period of time, and we believe that paper-based TOEFL should be offered as an alternative for at least a few years more.

A third use of tests in language programs is for placement purposes. *Placement tests* are often administered to find out the students' level of language acquisition and to incorporate him/her in a group that is suitable for his/her level. It has to be emphasized that decisions taken on the basis of these tests are of great importance to test takers as they affect them directly. It follows that students' further progress is a direct consequence of the truthfulness of their test scores. This means that if, for example, a student should be placed at an intermediate level but for some reason the test fails to show this, and the student is instead placed at an upper-intermediate or advanced level, the most viable result for the student would be frustration and failure to meet the objects of the program of study. An opposite case might also exist, one in which a student is placed at a level below his/her real level. This would undoubtedly result in the students' lack of motivation to continue, boredom and eventual dropping out of the program.

Finally, tests administered to students learning second or foreign languages aim to collect data as part of studies carried out to investigate various aspects of language learning and teaching. This might involve comparison of methods of instruction in terms of effectiveness, and in general the study of various variables that seem to affect language learning and which can be operationalized through test scores. Henning (1987) underlines that suitable tests need to be developed if we are to learn more about effective methods of teaching, strategies of learning, development and presentation of material for learning etc.

Test Usefulness

According to Bachman (1996), there are six qualities that make a test useful. They are; *reliability, construct validity, authenticity, interactiveness, impact* and *practicality*. He further notes that the traditional approach to defining these test qualities has been somewhat intuitive and has comprised a more or less separate description of them. In Bachman's view, test developers should try to strike the right balance between these qualities, which means that efforts to maximize them should aim towards maximising them all, not one at the expense of the other.

As a discussion of each and every aspect of test usefulness would require considerable time and space, this analysis will focus on the first two, that is reliability and construct validity. We will discuss three issues relevant to test reliability and validity.

1. What is test reliability and test validity?
2. Which are some threats to test reliability and test validity?
3. How can we provide evidence of reliability and validity?
4. Implications for teachers.

As these issues have a huge impact on the process of test design and development, teachers need to understand them and what is more they should try to apply this kind of knowledge when developing tests for their students.

Test Reliability

What is test reliability? Henning (1987) argues that a test is reliable when an examinee's test results are consistent if they take the same or a similar test more than once. According to Brown (1996) tests demonstrate reliability when, like other types of measuring instruments, they yield the same results every time measurement occurs. Bachman (1996) defines reliability as consistency of measurement.

For example, if a scale was used to weigh someone and it was found that that person weighs 78 kilos, we would expect that person to weigh 78 kilos again if he was weighed again after an hour. If, however, the scale shows 69, than we would think that there is something wrong with the scale and it would make us doubt its reliability.

If we administer the same test or two tests that are equivalent in terms of difficulty to the same group of students, in a time interval of no more than two weeks from each other, we should expect no significant difference between the means of the scores. In this case we would be able to claim that the test demonstrates reliability.

Test reliability is a function of the accuracy of the test. The more accurate a test is the more reliable it will be. However, unlike measuring heights, weights or distances, measuring language abilities is a much harder process as we are dealing with abstract notions which do not have referents in the objective reality.

Threats to Test Reliability

In order to be able to develop tests which allow teachers to infer correctly about their students' language ability, they need to be aware of the factors that influence test reliability. According to Brown (1996) these factors fall into three broad categories:

1. Environmental factors
2. Factors linked to administration procedures
3. Characteristics of the test items

Environmental Factors

There are a number of environmental factors that could have a negative impact on the students' performance. If a test is administered in a noisy, cramped environment where it is too hot or too cold, students' scores may suffer. Also, if the place is not properly lit, there is not enough space or the ventilation is bad, test reliability could also be negatively affected.

Other potential sources of test inconsistency could be psychological or physiological changes in test takers. (Henning, 1987). Factors such as physical or psychological fatigue, sickness, or any other emotional states might all lead to scores that do not reflect the students' real language ability. Even though these factors are often unpredictable and beyond teachers' control, conscious efforts should be made to maximize the conditions for the test to take place.

Factors linked to Administration Procedures

Factors that have to do with administrative procedures have also been found to contribute to the reduction of test reliability. According to Henning (1987), this occurs mostly when tests are administered to different groups of students in different locations or on different days. Factors such as unclear instructions, unsuitable time of test administration could all lead to a possible decrease in test reliability.

Characteristics of the Test Items

It has often been suggested that there is a relationship between test length and test reliability. The main argument to support this has been the one according to which longer tests do a better job of spreading students according to their level of ability than tests with few items in them. The level of test difficulty has also been reported as a potential factor affecting test reliability. Too difficult or too easy tests fail to separate students in terms of their ability. A third potential threat to test reliability linked to the nature of tests and test items has to do with the manner in which students respond to the examination. One is familiarity with test procedures. This means that when students have been exposed to the test format before, they seem to develop a certain kind of competence for guessing the right answer, which would normally lead to less reliable scores.

Ways of Providing Evidence of Reliability

There are several methods that can be applied to objectively check how reliable a test is. As some of them comprise complex statistical and mathematical formulas, our discussion will focus instead on two of them which we think could be applied by teachers in the process of developing tests for their teaching purposes.

Test-Retest Method

The same test is administered twice to the same group of students. The second administration happens no later than two weeks from the first administration. Students are not given feedback about the first administration and they are not warned about the second one. Students undergo no practice

in what they have been tested during this period. After the second administration, individual scores are arranged in two columns and some calculations are done. For example, group means are calculated and compared. Also individual scores on both administrations are compared with each other. If no significant differences are found, it could be claimed that the test seems to be reliable. As Brown (1996) maintains, although this method might sound a bit odd and provoke reactions when students are asked to write the same test twice, it could prove to be a practical way of finding out about the reliability of a test.

Parallel Tests

By this method two tests are administered to the same group of students. These tests are called parallel tests or equivalent tests because they are similar in terms of difficulty. The same procedures as above for the period after the test are applied. Now, although this method seems more natural than the test-retest method, it is harder because two alternatives of a test need to be written in which the criteria of equivalence have to be strictly met. It follows that the level of difficulty needs first to be defined and then test items need to be developed to match the difficulty level. This presents considerable difficulty for teachers first, due to the fact that they lack proper training in this field but also because of their heavy teaching load.

Implications for Teachers

It is clear that a test is as useful as it is reliable. Without reliable scores we would not be able to make objective inferences about test-takers' real level of knowledge acquisition. Although some factors that affect test reliability are clearly beyond the teachers' control (sickness, fatigue due to extraneous factors, etc), there are a number of things that teachers could do to maximize their chances of developing highly reliable tests for their students. Firstly, they should try to control for various environmental factors which might intervene in the normal course of test administration. For example, they should try to make sure that the time is right for the administration of the test. They should also try to make sure that the seating arrangement makes it possible to prevent cheating. Teachers should also strive that the test instructions are sufficient and clear.

The test is directly under the teacher's control. That is the reason why teachers should make every effort to produce tests that meet the criteria for maximizing test reliability and decreasing errors due to the nature of the test. First of all, the test should be reasonably long with no less than 75 items as Henning (1988) maintains. They should also be concerned with the level of test difficulty.

Test Validity

According to Hughes (1992), a test has validity when it matches the language skills or structures it intends to measure. For example, when designing a grammar test to help us identify to what extent students have acquired Present Simple or Present Continuous, we would set out to build a test or adopt one that revolves around these issues and not a test that focuses on other grammatical structures. Or, suppose we want to test our students' knowledge of vocabulary in chapter 4, which they have just covered. If for some reason vocabulary items for which our students have so far received no instruction were included in the test, we would definitely be reducing the validity of our test for it would clearly fail to match what it was designed to identify.

A very important concept relevant to test validity is *construct validity*. A construct is an abstract entity that does not have a referent in the objective reality. Language ability is an example of such a construct. In order to be able to measure this construct or ability, we need to define it. This means we need to have something that directly represents our construct and which is measurable or testable. Tests serve this purpose. For example, when we want to test students' ability to use Passive Voice, we would design a testing apparatus that deals with this grammatical structure in the hope that the scores we get will allow us to infer about our students' level of language ability. If for some reason items which are supposed to test Passive Voice include other structures, let's say Reported Speech or any other, the validity of the test would be seriously undermined.

In view of the above, it could be said that our test has construct validity to the extent that the test scores could be interpreted as indicators of the construct or ability we intended to measure (Bachman, 1996).

Factors that Affect Test Validity

A number of factors have been identified which seem to affect test validity in a negative way. Henning (1987) lists some of them. The main factor that affects test validity is the mismatch between a test and a construct it purports to measure. Another factor is an invalid application of tests. If, for example, a test which was designed to test vocabulary for first-year university students, was used with high-school students it would lack validity, although it might be quite valid for students for whom it was originally designed. A third factor that affects validity according to Henning (1988), is when standardized tests are developed on subjects from a distinct population, and these tests are administered to subjects from different population. TOEFL is such an example. The purpose of this test is that of testing foreign students who want to study in American universities. This test is designed based on foreign students of different ethnic or linguistic

backgrounds. If a university entrance examination of English language proficiency is sought for use with students applicants within a foreign country, all having the same native language, it follows that an exam of equal length would probably be more valid (Henning 1987).

Ways of Providing Evidence for Test Validity

How is test validity determined? Henning (1988) maintains that there are two main ways to determine test validity. One is through empirical methods. This involves the collection of data and the use of statistic formulae to calculate validity coefficients. The other is non-empirical based on inspection, intuition and common sense. As empirical methods need special training in statistics and the use of statistical computer programs to perform complex calculations, we would rather focus on non-empirical methods, given their practicality and the fact that they are well within the teachers' grasp and capability.

Although without empirical evidence it is somewhat difficult to claim in an objective way that a test is valid, through a number of practical actions teachers can maximize the chances of increasing the validity of their tests. It all starts with the definition of the construct or ability they want to test. For example, if they want to test their students' knowledge of grammar at the end of, let's say the elementary level, they need to think and be clear about what constitutes knowledge of grammar at the elementary level. In doing this they should draw on research that has been done on this field of language learning. Then, they should either write or adopt test items which directly match the grammar students have been exposed to during the elementary course.

Final Conclusion

Assessment in general and testing as one of the most important components of the assessment process constitutes a big challenge in language learning programs. Due to a total lack of training in test design and development, teachers rely wholly on their intuition and common sense or have recourse to test booklets that accompany the textbooks they use.

This, however, has a huge impact on students. Due to the fact that tests which are being used are often flawed, teachers fail to provide their students with objective feedback about their progress in foreign languages. This lack of knowledge affects teachers as well. Because they do not seem to be capable of assessing their students in an objective way, they cannot know what their students' strengths and weaknesses are. As a result they fail to adjust their course so that their students' weaknesses are addressed and strengths promoted (Brown 1996).

In view of the above, it is imperative that teachers should train in issues related to assessment and testing in particular. Also, another thing is that our educational institutions should start considering offering courses in test design and development along with other courses in foreign language methodology. This is something that has not been done before but these issues are too important to ignore any longer.

References:

Bachman, L. (1980). Fundamental Considerations in Language Testing. Oxford University Press.
Bachman, L. F., & Palmer, A. S. (1996). Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford University Press.
Brown, J. D. (1996). Testing in Language Programs. Prentice Hall Regents.
Grant, H. (1987). A Guide to Language Testing: Development, Evaluation, Research. Heinle & Heinle.
Hughes, A. (1992). Testing for Language Teachers. Cambridge University Press.